MSBD 5007 Homework 3

Pranav A, 20478966

October 30, 2017

Problem 1.

Let there be $x_1, x_2 \in S$, such that we get two inequalities, $Ax_1 \geq b$ and $Ax_2 \geq b$.

$$Ax_1 \geq b \tag{1}$$

$$Ax_2 \geq b \tag{2}$$

Multiplying, equation 1 by λ and equation 2 by $(1-\lambda)$, I get,

$$\lambda Ax_1 \geq \lambda b \tag{3}$$

$$(1-\lambda)Ax_2 \geq (1-\lambda)b \tag{4}$$

Adding eq. 3 and 4, I get,

$$\lambda Ax_1 + (1-\lambda)Ax_2 \geq \lambda b + (1-\lambda)b$$
$$\lambda Ax_1 + (1-\lambda)Ax_2 \geq b$$

Clearly, the above equation I get satisfies Jensen's inequality, which is the first requirement of convexity. Hence, given set S is convex.

Problem 2.1.

Given problem asks to find the proximity operator, $prox_{\beta_1||\cdot||_1 + \frac{\beta_2}{2}||\cdot||_2^2}$.
By definition,

$$prox_{\beta_1||\cdot||_1 + \frac{\beta_2}{2}||\cdot||_2^2}(y) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}||x-y||_2^2 + \beta_1||x||_1 + \frac{\beta_2}{2}||x||_2^2$$

By the condition of optimality, we have,

$$0 \in \nabla\left(\frac{1}{2}||x-y||_2^2\right) + \partial\left(\beta_1||x||_1\right) + \partial\left(\frac{\beta_2}{2}||x||_2^2\right)$$

Without the loss of generality, the objective function is separable. I will only consider i^{th} component here. This gives,

$$0 \in \nabla\left(\frac{1}{2}(x_i-y_i)^2\right) + \partial\left(\beta_1|x_i|\right) + \partial\left(\frac{\beta_2}{2}x_i^2\right)$$
$$0 \in x_i - y_i + \beta_1\partial|x_i| + \beta_2 x_i$$

Considering the case where $\partial(|x_i|) = 1$,

$$0 = x_i - y_i + \beta_1 + \beta_2 x_i$$
$$x_i = \frac{1}{\beta_2 + 1}(-\beta_1 + y_i)$$

This can happen when $y_i > \beta_1$, to make this zero.

Considering the case where $\partial(|x_i|) = -1$,

$$0 = x_i - y_i - \beta_1 + \beta_2 x_i$$
$$x_i = \frac{1}{\beta_2 + 1}(\beta_1 + y_i)$$

This can happen when $y_i < -\beta_1$, to make this zero.

Considering the case where $\partial(|x_i|) = [-1, 1]$, then x_i can assume the value within that domain to make it 0. Clearly, the domain of y_i here would be $y_i \leq |\beta_1|$

Putting together, I realize that it composes of proximal operators of $prox_{\beta_1 ||\cdot||}$ and $prox_{\beta_2 \frac{||\cdot||^2}{2}}$.

Thus,

$$prox_{\beta_1 ||\cdot||_1 + \frac{\beta_2}{2}||\cdot||_2^2}(y) = \frac{1}{\beta_2 + 1} prox_{\beta_1 ||\cdot||}(y)$$

where $prox_{\beta_1 ||\cdot||}(y) = T_{\beta_1}$, soft thresholding operator, conditioned on β_1.

Problem 2.2.

To apply the forward step to $\frac{1}{2}||Ax - b||^2$, the gradient would be taken, which is given by $\nabla \left(\frac{1}{2}||Ax - b||^2\right) = A^T(Ax - b)$.

To apply the backward step, proximal operator calculated in section 3.2 would be applied, which is $prox_{\beta_1 ||\cdot||_1 + \frac{\beta_2}{2}||\cdot||_2^2} = \frac{1}{\beta_2 + 1} T_{\beta_1}$.

Plugging these values in the proximal gradient update equation, I get the following algorithm which is given in Algorithm 1,

---

Algorithm 1 Elastic net proximal gradient

---

for $k = 1, 2, 3 \cdots$ **do**

$x^{(k+1)} \leftarrow prox_{\alpha_k(\beta_1 ||\cdot||_1 + \frac{\beta_2}{2}||\cdot||_2^2)}\left(x^{(k)} - \alpha_k \left(A^T \left(Ax^{(k)} - b\right)\right)\right)$

Problem 2.3.

Plugging the values in Nestrov's momentum algorithm I get the algorithm 2. The update rule and the preliminary assumption of setting the values have been shown below:

Algorithm 2 Faster elastic net proximal gradient

$x^{(0)} \leftarrow y^{(0)}$
$\lambda_0 \leftarrow 0$
for $k = 1, 2, 3 \cdots$ **do**
 $\lambda_k \leftarrow \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$
 $w_{k-1} \leftarrow \frac{\lambda_{k-1}-1}{\lambda_k}$
 $y^{(k)} \leftarrow x^{(k)} + w_{k-1}(x^{(k)} - x^{(k-1)})$
 $x^{(k+1)} \leftarrow prox_{\alpha_k(\beta_1||\cdot||_1+\frac{\beta_2}{2}||\cdot||_2^2)}\left(y^{(k)} - \alpha_k\left(A^T\left(Ay^{(k)} - b\right)\right)\right)$

Problem 3.1.

Given we have, $\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{c}^T\mathbf{x}$ with $\mathbf{Cx} \leq \mathbf{d}$ and $\mathbf{Ax} = \mathbf{b}$. Clearly, the lagrangian function is defined by,

$$L(\mathbf{x}, \lambda, \mu) = \mathbf{c}^T\mathbf{x} + \langle \lambda, \mathbf{Cx} - \mathbf{d}\rangle + \langle \mu, \mathbf{Ax} - \mathbf{b}\rangle$$

where $\lambda \in \mathbb{R}_+^p$ and $\mu \in \mathbb{R}^q$

It can be also written as,

$$L(\mathbf{x}, \lambda, \mu) = (c + \mathbf{C}^T\lambda + \mathbf{A}^T\mu)^T\mathbf{x} - \mu^T\mathbf{b} - \lambda^T\mathbf{d}$$

which is an affine function in $\mathbf{x}$. It follows that the dual function is given by

$$g(\lambda, \mu) = \inf_x L(\mathbf{x}, \lambda, \mu) = \begin{cases} -\mu^T\mathbf{b} - \lambda^T\mathbf{d} & \text{if } c + \mathbf{C}^T\lambda + \mathbf{A}^T\mu = 0, \\ -\infty & \text{otherwise} \end{cases}$$

Problem 3.2.

The dual problem is given by,

$$\begin{aligned} \max g(\lambda, \mu) \\ \text{subject to } \lambda \succeq 0 \end{aligned}$$

After making the implicit constraints explicit, we obtain,

$$\begin{aligned} \text{maximize} \quad & -\mu^T\mathbf{b} - \lambda^T\mathbf{d} \\ \text{subject to} \quad & c + \mathbf{C}^T\lambda + \mathbf{A}^T\mu = 0 \\ & \lambda \succeq 0 \end{aligned}$$

Problem 3.3.

Given function c^Tx is convex. If there exists a point $x \in \mathbb{R}^n$, which satisfies the conditions: $\mathbf{Cx} \leq \mathbf{d}$ and $\mathbf{Ax} = \mathbf{b}$, then Slater's condition is satisfied and the strong duality holds.

Problem 3.4.

Applying KKT condition here, I get,

1. $\frac{\partial L(\mathbf{x}^*, \lambda^*, \mu^*)}{\partial x} = 0, \implies c + \mathbf{C}^T\lambda + \mathbf{A}^T\mu = 0$
2. $g_i(x^*) \leq 0 \implies \mathbf{Cx}^* \leq \mathbf{d}$
3. $\lambda_i^* \geq 0 \implies \lambda \succeq 0$
4. $\lambda_i^* g_i(x^*) = 0 \implies \lambda_i(\mathbf{Cx}^* - \mathbf{d}) = 0$
5. $h_i(x^*) = 0 \implies \mathbf{Ax}^* = \mathbf{b}$

Problem 4.1.

For simplicity, I define $R(x)$ as $\sum_J ||x_{J_2}||_2$, which means taking Euclidean norm in the groups of 2 components each.

Thus when $x_{J_2} \neq 0$, $\partial R(x)$ would be simply $\frac{x_{J_2}}{||x_{J_2}||_2}$.

When $x_{J_2} = 0$, $\partial R(x)$ would be an element of $\{z : ||z|| \leq 1\}$

Problem 4.2.

Following the similar procedure like l1-norm proximity derivation, I get similar function to soft-thresholding,

$$prox_{\beta R}(y) = \frac{y_{J_2}}{||y_{J_2}||_2} \max(||y_{J_2}||_2 - \beta, 0)$$

Problem 4.3.

For a given step size α, one can have the following update rule,

Algorithm 3 Group lasso proximal gradient

for $k = 1, 2, 3 \cdots$ **do**
  $y^k = x^k - \alpha A^T(Ax^k - b)$
  $x_{J_2}^{k+1} = \frac{y_{J_2}^k}{||y_{J_2}^k||_2} \max(||y_{J_2}^k||_2 - \beta\alpha, 0)$

My solution can also work for keeping for other $J = 3, 4, 5, \ldots$.

Notes

1. This given solutions of the assignment follows the HKUST honour code. Although assignment has been discussed with other peers, the solutions are my own.
2. Kindly give feedback on this assignment on how to write up the solutions more elegantly. Your feedback is more important to us than grades.